## **GenAI Evaluation Tool: Metric Interpretation Logic Explained**

This document details the rationale and thresholds used for generating automated interpretations of metric scores within the GenAI Evaluation Tool's Streamlit interface. This applies to both the "Detailed Interpretation for a Single Test Case" and the "Interpreting Your Aggregated Results (Experimental)" sections.

**Guiding Principles for Interpretation:**

1. **Threshold-Based Categorization:** Most quantitative metrics are interpreted by comparing their scores against predefined thresholds. These thresholds typically classify performance into categories like "Strong/Good," "Moderate/Fair," or "Low/Poor." The specific values for these thresholds are heuristic and aim to provide a general sense of performance based on common expectations for these metrics. *They can and should be adjusted based on specific project requirements, domain nuances, and desired quality bars.*
2. **Contextual Feedback:** The interpretation provides qualitative feedback based on the score category and the nature of the metric. For instance, a low fluency score might suggest reviewing for grammatical errors, while a low factuality score would prompt a check for missing information.
3. **Actionable Suggestions:** Where appropriate, "Potential Actions" are suggested to guide the user in addressing identified weaknesses.
4. **Handling of Special Cases:**
   * **NaN (Not a Number):** Indicates that a metric could not be computed. This is often due to missing required input data (e.g., ref\_facts for Fact Presence), a calculation error, or because the metric is a placeholder. The interpretation will explicitly state this.
   * **Placeholder Metrics:** Metrics designated as placeholders (e.g., NLI Entailment, Professional Tone) are identified. The interpretation clarifies that they are not fully implemented and are expected to return NaN.
   * **Input Dependencies:** For metrics like Fact Presence and Key Point Coverage, the system first checks if the necessary input columns (ref\_facts, ref\_key\_points) were provided in the data. If not, the metric is deemed "Not Applicable" for that instance.
5. **Clarity on Scope:** For metrics like basic safety and PII checks, the interpretation emphasizes their rudimentary nature and advises manual review, especially if issues are flagged.

### **I. Logic for "Detailed Interpretation for a Single Test Case"**

This logic is primarily handled by the generate\_single\_case\_interpretation function in streamlit\_app.py.

**1. Fluency & Lexical Similarity (BLEU, ROUGE-1, ROUGE-2, ROUGE-L, METEOR)**

* **Rationale:** These metrics (scores 0-1) measure n-gram overlap (BLEU, ROUGE) or more advanced unigram matching with stemming/synonyms (METEOR) against a reference text. Higher scores mean more lexical similarity. An average of available scores provides a general fluency signal.
* **Logic:**
  + An average score is computed from all valid (non-NaN, non-placeholder) fluency metrics for the case.
  + **Thresholds (heuristic, for the average):**
    - >= 0.6: "Strong" (Suggests high overlap with reference text structure and wording).
    - >= 0.3 and < 0.6: "Moderate" (Some overlap, but noticeable differences).
    - < 0.3: "Low" (Significant dissimilarity in wording/phrasing).
* **Justification for Thresholds:** These are common sense starting points. Scores above 0.6 in metrics like ROUGE-L often indicate good summaries. Scores below 0.3 often point to substantial differences.

**2. Semantic Understanding (Semantic Similarity Score)**

* **Rationale:** This metric (score typically 0-1, but can be -1 to 1 for some sentence transformer models) assesses if the *meaning* of the generated text aligns with the reference, even with different wording.
* **Logic & Thresholds (heuristic, assuming scores are generally positive for similarity):**
  + >= 0.75: "Strong" (Meanings are very closely aligned).
  + >= 0.5 and < 0.75: "Moderate" (Core meaning likely captured, but nuances might differ).
  + >= 0.25 and < 0.5: "Fair" (Some conceptual overlap, but potentially missing key semantic elements).
  + < 0.25: "Low" (Meanings are likely quite different or unrelated).
* **Justification for Thresholds:** Based on typical behavior of cosine similarity for sentence embeddings. High similarity is usually >0.7. Scores <0.3 often indicate little to no semantic relationship.

**3. Trust & Factuality (Fact Presence Score)**

* **Rationale:** Checks for the explicit presence of critical, predefined factual statements in the answer. Score is the fraction of ref\_facts found (0-1).
* **Logic & Thresholds (heuristic):**
  + >= 0.7 (e.g., 70% of facts found): "Good."
  + >= 0.4 and < 0.7: "Moderate."
  + < 0.4: "Low."
* **Justification for Thresholds:** If many critical facts are provided, finding most of them (e.g., >70%) is good. Missing more than half is usually a concern.

**4. Completeness & Coverage (Key Point Coverage / completeness\_score)**

* **Rationale:** Checks if the answer addresses a list of expected broader topics or key points. Score is the fraction of ref\_key\_points covered (0-1).
* **Logic & Thresholds (heuristic):** Same as Fact Presence.
  + >= 0.7: "Good."
  + >= 0.4 and < 0.7: "Moderate."
  + < 0.4: "Low."
* **Justification for Thresholds:** Similar to Fact Presence, covering most key aspects is desirable.

**5. Classification Accuracy (Accuracy)**

* **Rationale:** For classification tasks, per-instance accuracy is binary.
* **Logic:**
  + Accuracy == 1.0: "Correct."
  + Accuracy < 1.0 (i.e., 0.0): "Incorrect."
* **Justification:** Straightforward; the prediction either matches the ground truth or it doesn't. (P, R, F1 are noted as 1.0/0.0 for the instance, with the caveat that their true meaning is in aggregation).

**6. Conciseness (Length Ratio)**

* **Rationale:** Compares the word count of the answer to the ground truth. Ideal ratio is often context-dependent, but extreme deviations are usually notable.
* **Logic & Thresholds (heuristic):**
  + < 0.5: "Significantly shorter." (Potentially too brief).
  + >= 0.5 and < 0.8: "Noticeably shorter."
  + >= 0.8 and <= 1.25: "Comparable length" (Often a good range).
  + > 1.25 and <= 1.75: "Noticeably longer."
  + > 1.75: "Significantly longer." (Potentially too verbose).
* **Justification:** A response half the length or nearly double the length of the reference often warrants a check. The "good" band around 1.0 (0.8 to 1.25) allows for some natural variation.

**7. Safety (Safety Keyword Score) & Privacy (PII Detection Score)**

* **Rationale:** These are basic, non-exhaustive checks. The goal is to flag *any* potential issue.
* **Logic:**
  + Score is 1.0 if no issues (keywords/patterns) are found, 0.0 if any issue is found.
  + **Observation is only generated if score is < 1.0** (i.e., an issue is detected). This avoids giving a false "all clear" for these basic checks.
* **Justification:** For safety/PII, even one instance of a problematic keyword or PII pattern is a concern that needs flagging.

**8. Placeholder Metrics**

* **Rationale:** To clearly communicate that these metrics are not yet functional.
* **Logic:** If is\_placeholder\_metric(metric\_key) is true and the score is NaN (as expected), a message is generated indicating its placeholder status and includes the metric's intended purpose from METRIC\_INFO.

### **II. Logic for "Interpreting Your Aggregated Results (Experimental)"**

This section provides interpretations for scores *averaged* across all test cases for a specific model and task.

**General Principles:**

1. **Focus on Non-Placeholder Metrics:** Interpretations are primarily generated for metrics that are fully implemented and have produced valid aggregated scores.
2. **Averaged Performance:** The logic applies thresholds similar to the single-case interpretation but to the *mean scores*. This gives a sense of the model's general tendency on that metric for the given task.
3. **NaN Handling:** If an aggregated score is NaN (e.g., fact\_presence\_score if ref\_facts were missing for all cases of a model/task), it's typically noted that the metric might not have been applicable or computable across the dataset.

**Specific Metric Interpretation Logic (Aggregated):**

* **Fluency & Lexical Similarity (Average of BLEU, ROUGE-L, METEOR):**
  + **Thresholds (heuristic, for the average):**
    - > 0.5: "Generally good fluency & lexical similarity."
    - > 0.2 and <= 0.5: "Moderate fluency/lexical similarity."
    - <= 0.2: "Low fluency/lexical similarity."
  + **Justification:** Slightly adjusted from single-case due to averaging; an average above 0.5 across many samples is generally decent.
* **Semantic Similarity Score (Average):**
  + **Thresholds (heuristic):**
    - > 0.7: "Good semantic similarity."
    - > 0.4 and <= 0.7: "Moderate semantic similarity." (Wider moderate band for averages)
    - <= 0.4: "Low semantic similarity."
  + **Justification:** An average semantic similarity above 0.7 is strong. Below 0.4 on average suggests consistent meaning divergence.
* **Fact Presence Score & Key Point Coverage (Average):**
  + **Thresholds (heuristic):**
    - > 0.7: "Good inclusion/coverage."
    - > 0.4 and <= 0.7: "Moderate inclusion/coverage."
    - <= 0.4: "Low inclusion/coverage."
  + **Justification:** Consistent with single-case; if, on average, less than 40% of required facts/points are covered, it's a systemic issue.
* **Classification (F1-Score or Accuracy if F1 is NaN - Average):**
  + **Thresholds (heuristic, for F1-score or Accuracy):**
    - > 0.75: "Good classification performance."
    - > 0.5 and <= 0.75: "Moderate classification performance."
    - <= 0.5: "Low classification performance."
  + **Justification:** F1/Accuracy above 75% is often considered good. Below 50% (random chance for binary) is poor.
* **Length Ratio (Average):**
  + **Thresholds (heuristic):**
    - 0.75 <= average\_ratio <= 1.25: "Good response length."
    - average\_ratio < 0.5: "Responses may be too short on average."
    - average\_ratio > 1.75: "Responses may be too verbose on average."
    - *Intermediate ranges are considered "acceptable" or "worth noting."*
  + **Justification:** Consistent average deviations from a 1:1 ratio are flagged.
* **Safety & PII Scores (Average):**
  + **Logic:** If the average score is < 1.0, it means at least one instance had an issue. The interpretation will flag this as a concern requiring manual review of individual cases, as averaging can mask isolated but critical failures.
  + **Justification:** For safety/PII, any failure is important. An average score doesn't negate individual breaches.

Important Caveat:

All thresholds and interpretation texts are heuristic starting points. They are designed to be generally useful but may not be perfectly suited for every specific use case, dataset, or quality standard. Users are encouraged to:

* **Review individual examples** that fall into different interpretation categories to understand if the thresholds align with their qualitative judgment.
* **Adjust thresholds** in the generate\_single\_case\_interpretation function or the aggregated interpretation logic within streamlit\_app.py if needed to better reflect their specific evaluation criteria and the expected performance profile of their LLMs.
* **Combine automated interpretations with human expertise.** These interpretations are aids, not definitive judgments.

This documentation should provide a solid backing for the choices made in the interpretation logic.